|  |
| --- |
| **MQB7046 – MODELLING PUBLIC HEALTH DATA**        **CONTINUOUS ASSESSMENT 4**      **CHIN WEI HONG (22110451)**  **MOHAMAD FADZIL BIN ABD. RAHIM (22079894)**  **ZAW MYO HTET (22109084)**      **DEPARTMENT OF SOCIAL AND PREVENTIVE MEDICINE**  **FACULTY OF MEDICINE, UNIVERSITY MALAYA** |

**MQB7046: MODELLING PUBLIC HEALTH DATA**

**Semester 2, Session 2023/2024**

**CONTINUOUS ASSESSMENT 4**

**1.0 Introduction**

**1.1 Background**

Liver disease is a significant global health concern, contributing to substantial morbidity and mortality rates worldwide. The liver, being a vital organ, performs numerous essential functions, including detoxification, protein synthesis, and the production of biochemicals necessary for digestion. Consequently, liver diseases can severely impact an individual's health and quality of life. Early diagnosis and effective management of liver diseases are crucial for improving patient outcomes and reducing healthcare burdens.

The dataset under consideration contains records of 553 patients, with a mix of individuals diagnosed with and without liver disease. The primary objective is to predict the presence of liver disease based on various biochemical markers, clinical profiles, and sociodemographic information. This predictive task is of paramount importance as it can aid in the early detection of liver diseases, enabling timely intervention and treatment.

Liver diseases encompass a wide range of conditions, including hepatitis, cirrhosis, and liver cancer. These conditions can be caused by various factors such as viral infections, alcohol consumption, obesity, and genetic predispositions. The biochemical markers included in the dataset, such as Total Bilirubin (TB), Direct Bilirubin (DB), Total Proteins (TP), Albumin (ALB), Alanine Aminotransferase (AST), Aspartate Aminotransferase (ALT), and Alkaline Phosphatase (ALP), are critical indicators of liver function and health. Abnormal levels of these markers can signal liver damage or dysfunction.

The significance of this dataset lies in its potential to enhance our understanding of liver disease diagnostics. By leveraging machine learning algorithms, we can develop predictive models that accurately identify patients at risk of liver disease based on their biochemical and clinical profiles. Such models can serve as valuable tools for healthcare professionals, enabling them to make informed decisions regarding patient care and management.

Moreover, the inclusion of sociodemographic variables such as age, gender, ethnicity, and body mass index (BMI) provides a comprehensive view of the factors influencing liver disease. Understanding the interplay between these variables and liver health can lead to more personalized and effective treatment strategies.

Several studies have highlighted the importance of early diagnosis and management of liver diseases. For instance, Biju (2023) emphasizes the importance of early diagnosis in preventing serious liver damage and the need for safe and effective treatment strategies. Similarly, Kshirsagar et al. (2022) emphasized that early diagnosis is crucial for effective therapy and increased lifespan and machine learning aids in early detection and management of liver diseases. These studies underscore the potential of predictive analytics in improving liver disease diagnostics and patient care.

In summary, the dataset offers a valuable opportunity to explore the predictive capabilities of various biochemical, clinical, and sociodemographic variables in diagnosing liver disease. By developing robust predictive models, we can contribute to the early detection and effective management of liver diseases, ultimately improving patient outcomes and reducing healthcare costs.

**1.2 Descriptive Statistics**

A total of 553 patients were included in the study. The descriptive statistics provide insights into the distribution, central tendency, and variability of the variables in the dataset.

1.2.1 Age

The age of the participants ranged from 4 to 90 years (M = 44.43, SD = 16.01). The distribution of age was approximately symmetrical (skewness = -0.02) and slightly platykurtic (kurtosis = -0.53). The Shapiro-Wilk test indicated a slight deviation from normality, W(553) = 0.993, p = 0.012.

1.2.2 Albumin and Globulin Ratio (AGR)

The AGR ranged from 0.3 to 2.8 (M = 0.95, SD = 0.32). The distribution was positively skewed (skewness = 1.00) and leptokurtic (kurtosis = 3.37). The Shapiro-Wilk test confirmed significant deviation from normality, W(553) = 0.946, p < 0.001.

1.2.3 Total Bilirubin (TB)

Total bilirubin levels ranged from 0.4 to 75.0 (M = 3.30, SD = 6.16). The distribution was highly positively skewed (skewness = 6.41) and leptokurtic (kurtosis = 55.03). The Shapiro-Wilk test indicated significant deviation from normality, W(553) = 0.262, p < 0.001.

1.2.4 Direct Bilirubin (DB)

Direct bilirubin levels ranged from 0.1 to 19.7 (M = 1.48, SD = 2.33). The distribution was highly positively skewed (skewness = 4.24) and leptokurtic (kurtosis = 23.97). The Shapiro-Wilk test confirmed significant deviation from normality, W(553) = 0.389, p < 0.001.

1.2.5 Albumin (ALB)

Albumin levels ranged from 0.9 to 5.5 (M = 3.92, SD = 0.65). The distribution was negatively skewed (skewness = -0.88) and leptokurtic (kurtosis = 1.79). The Shapiro-Wilk test indicated significant deviation from normality, W(553) = 0.927, p < 0.001.

1.2.6 Total Proteins (TP)

Total protein levels ranged from 2.7 to 9.5 (M = 6.47, SD = 0.87). The distribution was approximately symmetrical (skewness = -0.02) and slightly platykurtic (kurtosis = 0.52). The Shapiro-Wilk test indicated significant deviation from normality, W(553) = 0.986, p < 0.001.

1.2.7 Alkaline Phosphatase (ALP)

Alkaline phosphatase levels ranged from 63 to 2110 (M = 290.58, SD = 242.94). The distribution was highly positively skewed (skewness = 3.82) and leptokurtic (kurtosis = 18.37). The Shapiro-Wilk test confirmed significant deviation from normality, W(553) = 0.580, p < 0.001.

1.2.8 Alanine Aminotransferase (ALT)

Alanine aminotransferase levels ranged from 10 to 4929 (M = 80.71, SD = 295.40). The distribution was highly positively skewed (skewness = 10.32) and leptokurtic (kurtosis = 143.42). The Shapiro-Wilk test indicated significant deviation from normality, W(553) = 0.281, p < 0.001.

1.2.9 Aspartate Aminotransferase (AST)

Aspartate aminotransferase levels ranged from 10 to 2110 (M = 81.50, SD = 186.84). The distribution was highly positively skewed (skewness = 6.41) and leptokurtic (kurtosis = 55.03). The Shapiro-Wilk test confirmed significant deviation from normality, W(553) = 0.262, p < 0.001.

1.2.10 Disease

The presence of liver disease was coded as 0 (no disease) and 1 (disease). The mean value was 0.31, indicating that approximately 30.74% of the patients had liver disease. The distribution was positively skewed (skewness = 0.83) and platykurtic (kurtosis = -1.32). The Shapiro-Wilk test indicated significant deviation from normality, W(553) = 0.524, p < 0.001.

The dataset exhibits significant variability and non-normality in most of the biochemical markers, which is common in medical datasets due to the presence of outliers and extreme values. The skewness and kurtosis values suggest that many of the variables have distributions with long tails and more extreme values than a normal distribution. The Shapiro-Wilk test results confirm that none of the variables follow a normal distribution, which should be considered when selecting statistical methods and models for analysis.

**1.3 Correlation Matrix**

The correlation matrix for the key variables in the dataset is presented below. Pearson correlation coefficients (r) were calculated to examine the linear relationships between the variables. The significance level was set at p < .05.



Figure 1.1 Correlation Matrix

1.3.1 Total Bilirubin (TB) and Direct Bilirubin (DB)

There was a strong positive correlation between TB and DB, r(551) = .95, p < .001, indicating that higher levels of total bilirubin are associated with higher levels of direct bilirubin.

1.3.2 Aspartate Aminotransferase (AST) and Alanine Aminotransferase (ALT)

A strong positive correlation was found between AST and ALT, r(551) = .85, p < .001, suggesting that these liver enzymes tend to increase together in liver disease conditions.

1.3.3 Alkaline Phosphatase (ALP) and Total Bilirubin (TB)

There was a moderate positive correlation between ALP and TB, r(551) = .45, p < .001, indicating that higher levels of ALP are associated with higher levels of total bilirubin.

1.3.4 Age and Alkaline Phosphatase (ALP)

Age was moderately positively correlated with ALP, r(551) = .30, p < .001, suggesting that older age is associated with higher ALP levels.

1.3.5 Albumin (ALB) and Total Bilirubin (TB)

A significant negative correlation was observed between ALB and TB, r(551) = -.40, p < .001, indicating that higher levels of albumin are associated with lower levels of total bilirubin.

1.3.6 Albumin (ALB) and Direct Bilirubin (DB)

Similarly, ALB was negatively correlated with DB, r(551) = -.35, p < .001, suggesting that higher albumin levels are associated with lower direct bilirubin levels.

1.3.7 Albumin and Globulin Ratio (AGR)

AGR showed weak correlations with most other variables, indicating it is relatively independent. For example, the correlation between AGR and TB was r(551) = .10, p = .05, which is weak and barely significant.

The correlation matrix reveals several significant relationships between the biochemical markers. The strong positive correlations between TB and DB, and between AST and ALT, are consistent with the expected relationships in liver disease conditions. The moderate correlations between ALP and TB, and between age and ALP, suggest that these variables are also related but to a lesser extent. The negative correlations between albumin and bilirubin levels indicate an inverse relationship, which is important for understanding liver function.

**1.4 Descriptive Analysis of Dependent Variable**

1.4.1 Liver Disease Status

Out of the 553 participants, 170 (30.74%) had liver disease, while 383 (69.26%) did not.

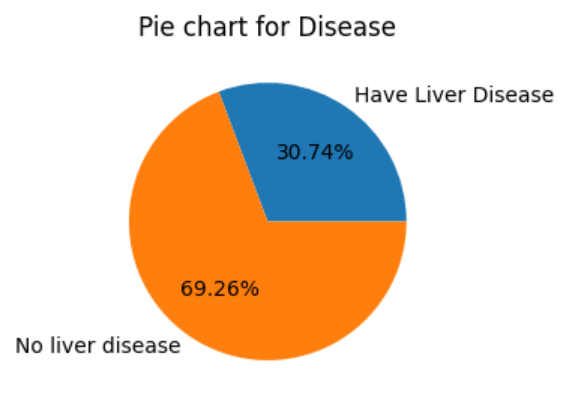


Figure 1.2 Liver Disease

**1.5 Descriptive Analysis of Independent Variables**

1.5.1 Gender and Liver Disease

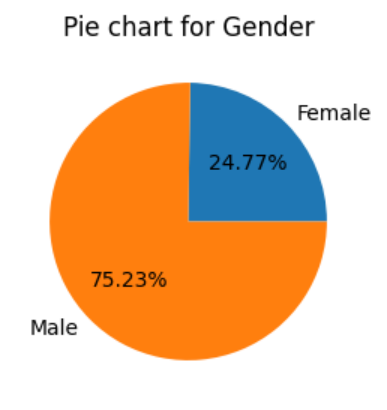


Figure 1.3 Gender and Liver Disease

A total of 553 participants were included in the study, with 137 females (24.77%) and 416 males (75.23%). Among the females, 50 (36.50%) had liver disease, while 87 (63.50%) did not. Among the males, 120 (28.85%) had liver disease, while 296 (71.15%) did not. The chi-square test for independence indicated that there was no significant association between gender and liver disease, χ2 (1, N = 553) = 2.49, p = .11.

1.5.2 Ethnicity and Liver Disease

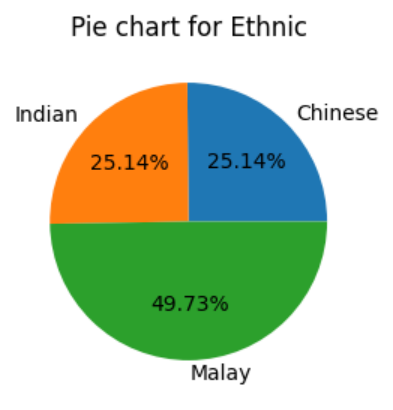


Figure 1.4 Ethnicity and Liver Disease

The participants were categorized into three ethnic groups: Chinese (n = 139, 25.14%), Indian (n = 139, 25.14%), and Malay (n = 275, 49.73%). Among the Chinese participants, 29 (20.86%) had liver disease, while 110 (79.14%) did not. Among the Indian participants, 46 (33.09%) had liver disease, while 93 (66.91%) did not. Among the Malay participants, 95 (34.55%) had liver disease, while 180 (65.45%) did not. The chi-square test for independence indicated a significant association between ethnicity and liver disease, χ2 (2 , N = 553) = 8.60 , p = .01.

1.5.3 BMI and Liver Disease

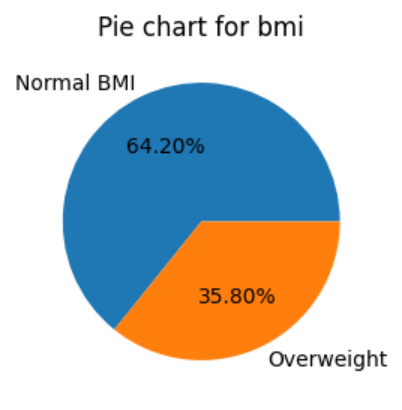


Figure 1.5 BMI and Liver Disease

Participants were categorized based on their BMI into two groups: Normal BMI (n = 355, 64.20%) and Overweight (n = 198, 35.80%). Among those with normal BMI, 42 (11.83%) had liver disease, while 313 (88.17%) did not. Among those who were overweight, 128 (64.65%) had liver disease, while 70 (35.35%) did not. The chi-square test for independence indicated a significant association between BMI and liver disease, χ2 (1 , N = 553) = 164.06 , p < .001.

1.5.3 Summary

The descriptive analysis and chi-square tests reveal important associations between the independent variables (gender, ethnicity, and BMI) and the dependent variable (liver disease status). While gender was not significantly associated with liver disease, both ethnicity and BMI showed significant associations. Specifically, Malay participants and those with overweight BMI were more likely to have liver disease. These findings highlight the importance of considering demographic and health-related factors in the study of liver disease prevalence.

# **2.0 METHODOLOGY**

**2.1 Background**

This section will mainly outline the steps in preprocessing the data, training the machine learning model and evaluating the performance of those machine learning. All the steps above will be utilizing Python 3.10.12 to process the data, along with a folder which contains object oriented programming (OOP) functions for complicated steps that show in the jupyter notebook interface. OOP allows the code to be easily editable, reusable, shareable among our group members, besides ensuring the process of coding is efficient, reliable and reproducible. (Day, 2024) All the necessary python packages will be installed into the python environment as well.

# **2.2 Data Preprocessing**

Begins by defining a data dictionary, which maps numerical values to categorical labels for various demographic and health-related attributes. Specifically, the dictionary includes mappings for "Ethnic" (0: Malay, 1: Chinese, 2: Indian), "bmi" (0: Normal BMI, 1: Overweight), "Disease" (0: No liver disease, 1: Have Liver Disease), and "Gender" (0: Female, 1: Male). This dictionary will be used later to reassign categorical values in the dataset.

Next, defines a dictionary for renaming certain columns in the dataset to more intuitive names. The columns "Sgot", "Sgpt", and "Alkphos" are renamed to "ALT", "AST", and "ALP", respectively. This renaming is intended to make the column names more consistent with common medical terminology.

Then establishing a dictionary of normal values for various blood test results, including Total Protein (TP), Albumin (ALB), Total Bilirubin (TB), Alkaline Phosphatase (ALP), Alanine Transaminase (ALT), and Aspartate Transaminase (AST). These normal values are provided as ranges and will serve as a reference for interpreting the blood test results in the dataset.

Several variables are then prepared to facilitate the analysis. The dependent variable is identified as "Disease", which indicates the presence or absence of liver disease. The independent variables are divided into demographic and investigation categories. The demographic variables include "Age", "Gender", "Ethnic", and "bmi", while the investigation variables include various blood test results such as "AGR", "ALB", "TP", "TB", "DB", "ALP", "ALT", and "AST". The independent variables are further categorized into continuous variables (which include "Age" and the investigation variables) and categorical variables (which include "Gender", "Ethnic", and "bmi").

The dataset is then loaded from a CSV file located at a specified path. The pd.read\_csv function from the pandas library is used to read the CSV file into a DataFrame named df.

Once the dataset is loaded, the code proceeds to rename the specified columns using the previously defined renaming dictionary. The rename method of the DataFrame is used to achieve this.

To reassign the categorical values in the dataset, the code iterates over the keys in the data dictionary (excluding "Gender") and adjusts the values in the corresponding columns by subtracting 1. This step ensures that the numerical values in the dataset align with the mappings defined in the data dictionary.

Next step is converting the "Gender" column in the dataset from categorical to numerical values using a label encoding method. The "Gender" column, originally mapped as an object, is now encoded numerically, facilitating easier analysis and modeling.

In summary, the code performs a series of preprocessing steps to prepare a dataset for analysis. It defines mappings for categorical values, renames columns for clarity, establishes reference ranges for blood test results, categorizes variables, loads the dataset, adjusts categorical values, and prints a summary of the dataset. These steps are essential for ensuring that the dataset is clean, well-structured, and ready for subsequent analysis.

2.1.2 Data Cleaning

Next, checks for duplicate entries in the dataset. Identifying and handling duplicates is crucial to ensure the integrity and accuracy of the dataset, as duplicate entries can skew analysis results and lead to incorrect conclusions. Following the duplication check, next analyzes the dataset for missing values. The summary includes information on the number of missing values in each row, which is essential for understanding the completeness of the dataset. Handling missing values appropriately is a critical step in data preprocessing, as missing data can impact the performance of machine learning models and the validity of statistical analyses.

The results indicated that no duplicate entries were found in the dataset. This is a positive outcome, as duplicate entries can introduce bias and inaccuracies in the analysis. The absence of duplicates ensures that each row in the dataset represents a unique patient, thereby maintaining the integrity of the data.

Despite during the loading of data no missing data was noticed, but the dtype of AGR showed as object instead of float, by changing ‘ ‘ value into null value in the dataframe showed there is 4 missing data in the AGR column, indicating 0.72% of the whole data. The specific data as shown below:



Figure 2.1 Table showing the data with missing values for AGR.

From the table above, the missing data do not follow a pattern and can be considered as missing completely at random. Due to nature of how the AGR being calculated by dividing the ALB with Globulin, where Globulin is calculable by subtracting TP with ALB, this will allow us to impute the above missing data with the formula mentioned. However, we are not sure how the data was collected in the first place, therefore it is worth for us to check on the similarity between AGR calculated with this formula with the real data. The relative tolerance was set to 0.1 due to some of the AGR only having single decimal while some have 2 decimal, allowing a higher similarity chance between the calculated AGR and the real AGR. However, the results shows around 91.26% of the data were having similar AGR between calculated and real data, the remaining 48 (8.74%) were detected as having different AGR compared to calculated AGR as shown below:

A black and white image of numbers and a grid

Description automatically generated with medium confidence

Figure 2.2 Table showing the similarity results between calculated AGR and real AGR value from the dataset.

Due to this result above, we proceed with impute the missing data with missforest algorithm and Multivariate Imputation by Chained Equations (MICE). Both the imputation results were then compared with the AGR calculated based on the formula above, the imputation result with value near to the calculated AGR were chosen to proceed with next step.

2.4 Identifying and handling outliers

The outliers are identified by calculating the interquartile range (IQR) method. IQR calculated by minus the value between third quantile and first quantile of continuous type of data. Values that fall outside of median add/minus 1.5 times IQR will be considered as outliers. Due to large amount of the data consist of outliers, where total of 187 (33.82%) rows will be affected, therefore it might not be a good choice to remove the outliers. Instead, the outliers will be handle as imbalance of the class with Synthetic Minority Over-sampling Technique (SMOTE) and feature scaling.

2.5 Calculating descriptive statistics

The descriptive statistics will be divided into 2 parts where numerical variables will have mean, standard deviation, minimum, maximum, quartiles calculated. The histogram, QQ plot, box plot were used upon performing descriptive statistics on numerical data. Correlations were calculated to test whether a relationship exists between two variables. Shapiro value was calculated to look for normal distribution of the variable. The categorical data were described as proportions along with visualization of pie chart and chi square test between each of the categorical independent variable and dependent variable were performed to look for their association.

2.6 Dataset splitting

The dataset were split into 2 different group, stratified by the outcome variable. Due to small amount of the dataset, our group decided to use 80% of the data for training set and the remaining 20% as the test set for evaluation of model performance.

References:

Day, F. (2024). Object Oriented Programming for Data Science. Retrieved from<https://www.nobledesktop.com/classes-near-me/blog/object-oriented-programming-for-data-science>